university of groningen

campus fryslân

Synthesizing Anger: Enhancing Emotional Speech from Text in Novel Dialogues

Jocomin Thomas-Luc Michel Galarneau

11/06/2024

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Synthesizing Anger: Enhancing Emotional Speech from Text in Novel Dialogues

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Shekhar Nayak (Voice Technology, University of Groningen)

Jocomin Thomas-Luc Michel Galarneau (S3291561)

June 11, 2024

Acknowledgments

I want to thank my supervisor and lecturers for guiding me onto a clear path with this research and for pushing me to put in my maximum effort to make the most of this programme.

Thank you to my wonderful classmates, which without our strong bonds created through stress, spite, and tears, I would never had made it this far.

I acknowledge the Center for Information Technology of the University of Groningen for their technical support and for providing access to the Hábrók high-performance computing cluster.

To my family, who supports me with all their love from across the pond and continuously encourage me to forge my own way.

And a deep personal love and appreciation to my wonderful partner who acted as my soundboard, therapist, partner-in-crime, and source of my joy and safety. Without you, who knows where I'd be? Probably making ocean sounds by myself.

Lastly, I would like to express my gratitude to all the individuals who have played a part, no matter how small, in shaping my academic journey and the successful completion of this thesis.

Abstract

This thesis aims to enhance the synthesis of emotional speech for text, focusing on anger as portrayed in novel dialogues. Building upon advancements in emotion recognition and speech synthesis technologies, the research maps textual emotion descriptors to dimensional parameters of arousal and valence to authentically synthesis the nuances of anger. Using Ekman and Cordaro (2011) basic emotions, Russell (1980) circumplex model, and the Expressive-FastSpeech2 speech synthesis model (K. Lee, 2021), the study intends to generate speech that faithfully represents varying levels of anger, based on novel dialogue contexts and sentiments. In an experimental setup, 12 anger-labeled and 8 neutral-labeled lines from "Alice in Wonderland" (Carroll, 2006) were synthesized into groups with different intensity levels. Twenty participants were then asked to judge up to two emotions and their intensities within the synthesized samples on a scale from 1 to 7. Results indicate a 55.25% participant recognition rate for the intended emotion within the audio sample, with a 32.65% rate for intensity levels, considering scores that fell within one value of the correct level. While this thesis does not conclude that the current methodology consistently creates authentic nuanced emotions, it reaffirms the importance of textual context for novel dialogue, highlights the variability in individual emotional perceptions, and provides a groundwork for future studies to refine and build upon.

Keywords: emotions, sentiment recognition, arousal, valence, expressive speech synthesis

Contents

1 Introduction

Audiobooks draw a niche transition between text and media, allowing us to listen to a reader's interpretation of the literature, filled with their own acting choices in emotion, accent, and cadence, drawing listeners into the setting of the story that much more. However, interpretations of literature are not unanimously agreed upon and, even more so, contradictions between two different works of art on the same piece that were created from "faulty" interpretations are widely scorned. A striking example is the divergent portrayal of Albus Dumbledore's reaction in "Harry Potter and the Goblet of Fire", where the film's aggressive tone of Dumbledore "crashes open the door" to speak to Harry starkly contrasts with the book's calm demeanour, where Dumbledore "asked calmly" (Kloves & Audsley, 2005; Rowling, 2000).

It came to our interest, then, that voice technology might have a place in addressing this issue of contradictory interpretations, by providing a more accurate rendition, close to the original version. Emotion recognition and text-to-speech (TTS) models have continuously progressed over the years, drawing closer in synthesizing natural speech and refining human-machine interaction. However, many of these studies have not attempted a generalized approach to synthesizing expressive speech, which might better serve as a basis for future interpretations of the work. While the fear of such a program replacing audiobook actors is evident, the intention behind this study is to create a quick and efficient method to synthesize novels and other forms of literature, in order to get a basic idea of how a character might sound like in a specific situation.

There are a couple points of thought to this study that must be made clear. In text and literature, the emotion, or sentiment, of dialogue is created through either surrounding context in exposition or dialogue descriptors, such as "asked calmly" in the example above. Therefore, emotional recognition via text is not challenging, as long as certain keywords are placed near the dialogues, classifying it with specific emotions. However, natural speech does not have those descriptors, as they are produced by soundwaves that have their own acoustic features such pitch and loudness that fluctuate to form emotions, which literature, of course, does not have. Many expressive-speech synthesis models have been able to overcome this obstacle in a multitude of ways, such as implementing acoustic features into code, in order to create basic emotional speech. Yet, literature is not only made up of basic emotions, instead imparting a variety of flavor and intensity to these dialogues, making them rich in prose. Thus, a middle-ground must be found, which comes in the form of dimensional parameters such as arousal and valence, which oversee the vastness of emotional range, and can be used to synthesize more nuanced emotions. The goal of this study is to determine if the levels of intensity of chosen base emotions can be used as a generalization for synthesizing more varied expressive speech, to eventually create a basis for novel character interpretations.

Now that a brief motivation for this research has been presented, the structure of the thesis is the following: subsection 1.1 introduces the research question posed along with a hypothesis on the outcome of the research. Section 2 provides an extensive literature review that frames the research question and hypothesis in the state-of-the-art, and explains the concepts and theories that the study incorporates. In section 3, the methodology is covered and the underlying models used are explained, then the experimental setup developed to answer the research questions is outlined. Section 4, describes and and discusses the obtained results in detail and how they affect the research question and hypothesis, while also noting limitations to the study. Lastly, section 5 summarizes the thesis and presents the conclusions drawn, along with recommended future work.

1.1 Research Question and Hypothesis

To best summarize the previous discussion, our research questions can be formulated as such:

> **To what extent can we translate emotion classification scores to variable arousal and valence scores in order to synthesize nuanced emotional speech?**

This can be broken down into further sub questions:

- Can the coded emotions be consistently recognized in the nuanced synthesis?
- Is the level of intensity of the emotion of the synthesis consistently and correctly perceived?
- How reliable is the creation of a generalized basis for the transfer of emotion classification scores to appropriate arousal and valence scores of synthesized emotion?
- What are the challenges involved in the process?

Our hypothesis is that (H1) a generalized basis for sentiment text classification to emotional speech features can be achieved, resulting in (H2) synthesized intensity levels that will be recognizable, within a small range, and (H3) that coded base emotions can be found consistently. The falsification could indicate that generalization with the chosen models does not work, that the dataset and synthesis model chosen were not proper for this study, or that humans may not be able to consistently pick out base emotions from a more nuanced emotion or recognize generalized nuanced differences in emotional states.

2 Literature Review

Emotions and their machinations have been a long-researched topic, and the answers that they hold to best define them has been sought after by various fields. The ability to recognize and simulate emotion via a machine is a cornerstone in human-machine interactions and natural language processing and has resulted in a variety of innovative models and theories. This section is dedicated to related works and concepts that were researched, considered, and incorporated when formulating this study. The review is split into three sections, exploring the existing research on each of these interconnected aspects of emotion and sentiment analysis: Emotions and the approaches to classifying them, Emotion and sentiment recognition via text and audio, and Emotional speech synthesis.

Subsection 2.1 covers the concepts of emotions and the evolution of the definitions surrounding them, with a focus on two fields of study: perception-oriented studies and acoustic-oriented studies. By illustrating the surrounding theories of emotions and their features, this subsection sets the stage for understanding the scope and difficulties surrounding emotional recognition and synthesis.

Subsection 2.2 explores the multitudes of models and concepts of emotional recognition in audio and text. We will turn our attention to text-based recognition models and theories, touching upon three different approaches of recognition: keyword-based, lexicon-based, and machine learning-based. The model chosen for this study, and research it derives from, will be explored.

Subsection 2.3 entails the concepts and approaches of emotional synthesis. Emotional synthesis carries its own issues with it but has pushed forward multiple answers to combat these problems. Models that respond to these obstacles will be explored, eventually deliberating on the motivation for chosen TTS model of this study, outlining the advantages it has over the other models. This subsection hopes to illuminate the evolving field of emotional synthesis and give insight into the future of human-machine interaction.

Through an in-depth analysis of these three thematic areas, this literature review endeavors to provide a comprehensive overview of the current state of research in emotion analysis and synthesis. By synthesizing insights from diverse disciplines and methodologies, it seeks to identify key challenges, emerging trends, and opportunities for future research in this vibrant and interdisciplinary field.

2.1 Emotional Concepts and Theories

The concept of emotion and an accurate way to describe it has eluded multiple fields of study for decades, ranging from biology and psychology to communication and machine learning. Though no singular definition has been accepted by all branches of researchers, several theoretical approaches have brought forth extensive explanations, while insight into speech and linguistic features have assigned values to "basic" forms of emotion. In this section, we explore multi-disciplinary concepts and attributes of emotion, in order to gain better insight into the underlying task that this study is attempting to replicate.

Understanding emotion and the complexity of its workings has been an uphill battle for researchers and the average person for generations, though recent undertakings in a myriad of fields have allowed us to grow closer to being able to explain them. Overall, two general approaches have stood the test of time and are used consistently by researchers (dimensional and discrete approaches), while a third has slowly fallen into obscurity (prototype approach) (Guerrero, Andersen, & Trost, 1996; Harmon-Jones, Harmon-Jones, & Summerell, 2017). The dimensional approach to emotion identifies emotions based on their placement on dimensions, which can come in the form of valence (positive vs. negative affect/feeling), arousal/activity (calm vs. excited), intensity (strong vs. weak) (Guerrero et al., 1996), or motivational direction (inclination to approach or avoid something) (Harmon-Jones et al., 2017). These dimensions intersect to classify the multitude of emotions and feelings that can occur in humans, and at times animals, allowing for an acute overview of the emotional range. The discrete approach, however, posits that a small number of emotions are distinct from one another, and are so instinctively built within humans, that they can be easily recognized and distinguished cross-culturally. This approach further states that emotions other than these "instinctual" emotions, are either delineations of the universal emotions or are "blends" of two or more of these primary emotions (Guerrero et al., 1996). The third approach, the prototype approach, has become less popular as the years past, despite its original promise of finding common ground between the discrete and dimensional approaches. The prototype approach creates "families" of emotions that are distinguished via characteristics, such as valence of function, then creates clusters within those families that are separated by characteristics such as intensity. The caveat to this approach is that it is built upon people's experiences that provide them with the information to conceptualize and categorize these emotions. However, the approach has been met with multiple criticisms, such as that people's accounts of emotional experiences may be insufficient to describe complex emotions, the continuous disagreement over which emotion qualify as being "basic", and that some emotions may cross the preconceived boundaries, for instance excluding the idea of emotional blending (Guerrero et al., 1996). Perhaps because of these issues, research related to prototype approaches to emotion have been few to none, as evident by the publication year of the literature it was mentioned in. However, the two main approaches that we will be covering are nonetheless met with criticism, and as will be made evident, the rise for hybrid models incorporating elements of both approaches is not as surprising.

As was described before, the discrete approach states that there are innate, universally recognized emotions, which have a fixed set of neural and bodily expressed components, are distinguishable from one another, and have a fixed feeling or motivational component that has been selected through extensive interactions with ecological stimuli (Tracy & Randles, 2011). While these features have been widely agreed upon by researchers, the actual choice of which emotions fit all these features and can be described as such is still a point of contention to this day. Paul Ekman, who pioneered

the rise of the discrete approach, believed strongly in the cultural impact and relevance to basic emotions, contending that while innate, these emotions are partly dependent on cultural influence and are inherent in the learning process during the developmental process (Ekman & Cordaro, 2011; Guerrero et al., 1996; Harmon-Jones et al., 2017; Tracy & Randles, 2011). As such, he stipulated several characteristics found in nearly all basic emotions, seen in Figure 1.

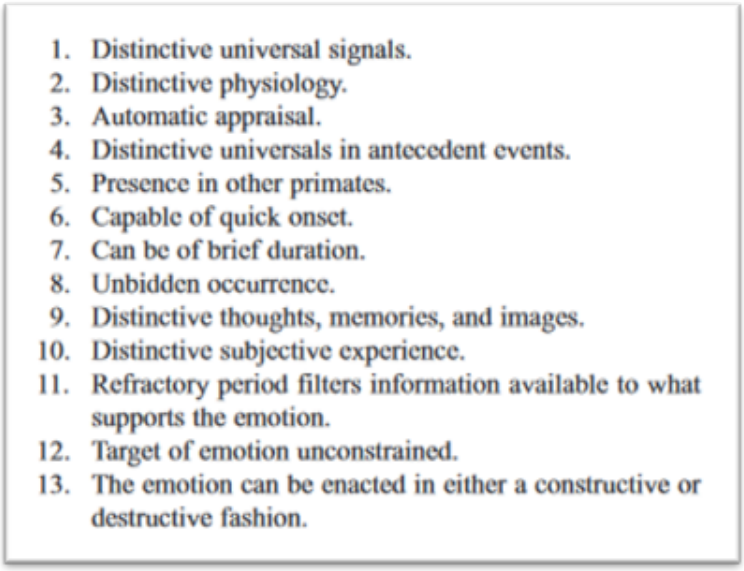
1. Distinctive universal signals.
2. Distinctive physiology.
3. Automatic appraisal.
4. Distinctive universals in antecedent events.
5. Presence in other primates.
6. Capable of quick onset.
7. Can be of brief duration.
8. Unbidden occurrence.
9. Distinctive thoughts, memories, and images.
10. Distinctive subjective experience.
11. Refractory period filters information available to what supports the emotion.
12. Target of emotion unconstrained.
13. The emotion can be enacted in either a constructive or destructive fashion.

Figure 1: Paul Ekman and Cordaro (2011) characteristic of basic emotions

Interestingly, Ekman and Cordaro (2011) rejected the idea that emotions such as anger and disgust were inherently destructive or negative, and believed that any emotion had the capability to be committed in an act of destruction or construction. As such, he pushed six emotions, now seven, as "basic" emotions (Ekman & Cordaro, 2011; Guerrero et al., 1996): [1] Anger: the response to interference with our pursuit of a goal we care about. [2] Fear: the response to the threat of harm, physical or psychological. Often triggers anger. [3] Surprise: the response to a sudden unexpected event. Very brief. [4] Sadness: the response to the loss of an object or person to which you are very attached. Can fluctuate. [5] Disgust: repulsion by the sight, smell, or taste of something. Can also be provoked by people's actions or ideas. [6] Contempt: feeling morally superior to another person. [7] Happiness: feelings that are enjoyed, that are sought by the person. Variable trigger and behavior. These seven emotions can encompass a wide range of behavioral reactions to stimuli, each distinct from one another, but also allowing for the possibility of emotional blending. Ekman and Cordaro (2011) continue, expressing that these basic emotions are rarely elicited in a pure form, changing rapidly, and blending into one another as the environment changes, something that other researchers have agreed on (Tracy & Randles, 2011). Furthermore, he posits that each emotion, when adapting to deal with life tasks, resulted in physiological changes as well, stating the increase in evidence that specific autonomic nervous system (ANS) responses correspond to specific basic emotions (i.e. increase blood flow to arms and hands for anger, gag reflex for disgust). Ekman's discrete approach and basic emotions have built a basis for others to contend and evolve their own theorems from, while also retaining the credibility and simplicity to be used above other concepts. While many have agreed with Ekman's findings, several researchers from a variety of disciplines

have gone further, adding more emotions to the list, and arguing a difference in the characteristics listed, which is shown in Figure 2.

Theoretically and empirically supported basic emotions according to each model			
IZARD	PANKSEPP & WATT	LEVENSON	EKMAN & CORDARO
Happiness	PLAY	Enjoyment	Happiness
Sadness	PANIC/GRIEF	Sadness	Sadness
Fear	FEAR	Fear	Fear
Anger	RAGE	Anger	Anger
Disgust		Disgust	Disgust
Interest	SEEKING	Interest?	
Contempt?			Contempt
	LUST	Love?	
	CARE	Relief?	Surprise

Figure 2: Side by side comparison of what different researchers classify as "Basic Emotions" (Tracy & Randles, 2011)

Levenson (2011) agrees with a majority of Ekman's theories yet maintains a slightly larger list and holds that his criteria are more general and rooted in neurological aspects. He quickly explains his list of basic emotions, stating much of the same as Ekman, including enjoyment (happiness), anger, disgust, fear, surprise, and sadness, but also adds relief/contentment, interest, and love. However, he elaborates and states that while strong evidence exists for the first six emotions, existing evidence for the distinctness and hard-wiredness/continuity of the latter three emotions is lacking. While following with the same groupings of criteria as Ekman (distinctness, continuity, and function), Levenson (2011) focuses in particular on the neurological aspects of behavior, in this case the stimulation/blocking of particular brain circuits to see when particular emotions appear or disappear.

Continuing in the field of neuroscience, Panksepp and Watt (2011) create their own list of basic emotions, denoting it as "prototype emotional states", emotions that can be evoked by artificial activation of subcortical networks of the brain. While they are described as "states", lines of connection can be drawn between these states and the basic emotions. Their list includes "SEEKING, FEAR, RAGE, LUST, CARE, PANIC/GRIEF, and PLAY" (Panksepp & Watt, 2011). PLAY, GRIEF, FEAR, and RAGE can easily be translated to happiness, sadness, fear, and anger respectfully, while SEEKING comes in line with Levenson (2011) and Izard (2011) own beliefs in adding 'interest' to the list. Yet, at the fault of neuroscience, Panksepp and Watt (2011) dismiss disgust as an emotional state, as they believed it evolved to help regulate physiological needs. However, as Tracy and Randles (2011) point out, the fact that disgust "influences behavior in response to [...] specific stimuli [...] encountered in humans' ancestral environment, has a distinct and cross-culturally recognized nonverbal expression, dedicated neural circuitry, and interact with cultural learning to produce emotional schemas", indicates that disgust is not limited to only physiological states.

Contrastively, Izard (2011) does not consider himself a 'basic emotion theorist', noting that rather than focusing on 'basic emotions' he sets his attention on emotion schemas, which he believes to be the emotions of everyday life. Of the basic emotions that he has identified, Izard follows the same emotions as Ekman, with two points of contention. Like Levenson (2011), he adds interest as a basic emotion, though affirms that there is sufficient evidence to support its choice according to

his own criteria. He is, however, reluctant to label contempt as a basic emotion, believing it to be too much of a learned emotion rather than a "basic" emotion, sharing the belief of Levenson (2011) that basic emotions must be innate. Furthermore, all three researchers argue that while learning and experience may shape the circuitry for basic emotions, they still must be innate and built into the nervous system, rather than be learned de novo, or from the beginning, something that puts them at odds with Ekman and Cordaro (2011). The four researchers state that a basic emotion must be evolutionary shaped and biologically prewired, and that their underlying function to solve problems during evolution are what differentiates them from each other, though each have certain caveats to the statement. Tracy and Randles (2011) note that more studies are being published with neurological underlinings, something that is noted by the researchers they interviewed as being instrumental in understanding "basic" emotions.

However, while this bodes good news for the fields of psychology and neuroscience, how does this impact the field of AI and machine learning? While the architecture of an AI model includes "neural networks" it still fundamentally works differently than an organic brain. Indeed, the basic emotions outlined by Ekman and Cordaro (2011) can be used to classify sentiment and emotion more easily in recognition software, but asking models to pick up on changes in emotions without the capabilities to watch physical expressions or how to best replicate them limits the avenues in which a model would be able to complete this task. One possible solution may come in the form of the other approach to emotion classification, dimensional.

The dimensional approach differs from the discrete approach in that it is not limited to a set of "basic emotions" from which every other sentiment is derived from. Instead, it uses multiple dimensions to group emotions together based on their relative positions to the extremes of the chosen dimensions. As mentioned before, the two main dimensions that are used in nearly all forms of models utilizing the dimensional approach, are valence (positive vs. negative) and arousal (calm vs. excited) (Harmon-Jones et al., 2017; Laukka, Juslin, & Bresin, 2005). Although the conceptual reason for focusing on valence and arousal is sound, allowing for more wide-spread organization, a more analytical reason why the two attributes are highly lauded can be found in their correlation and association with mean fundamental frequency, F0 variability, speech rate, voice intensity (loudness), and frequency energy (Laukka et al., 2005). Starting with these two base dimensions, Russell (1980) created the circumplex model, a two-dimension, circular structure that was cut in four quadrants and plotted emotions based on their level of arousal and valence (Guerrero et al., 1996). The clear structure allows for easy categorization for the multitude of emotions that an individual may feel and allow for a more nuanced approach than is available for the discrete approach. As shown in Figure 3, which is a more completed circumplex model used for emotional classification in video games (Zagalo, Torres, & Branco, 2005), the "main" or "basic" emotions of the discrete approach can be easily plotted separately, and delineations can find a place near the base emotions to show connections as well as inverse relationships.

However, the circumplex model has been heavily criticized, with multiple researchers pointing out that two dimensions is not enough to capture the complexity of emotions, noting its simplicity in design (Guerrero et al., 1996; Harmon-Jones et al., 2017; Laukka et al., 2005).

To answer this issue, researchers began adding more dimensions and attributes, beginning with Daly, Lancee, and Polivy (1983), who added emotional intensity (strong vs. weak) in order to further differentiate emotions and their placement with the blocks. For example, as seen in Figure 2, contempt and agitated are quite close together, indicated that they must be correlated quite closely together. However, realistically speaking, the two emotions, when felt, are quite different. With

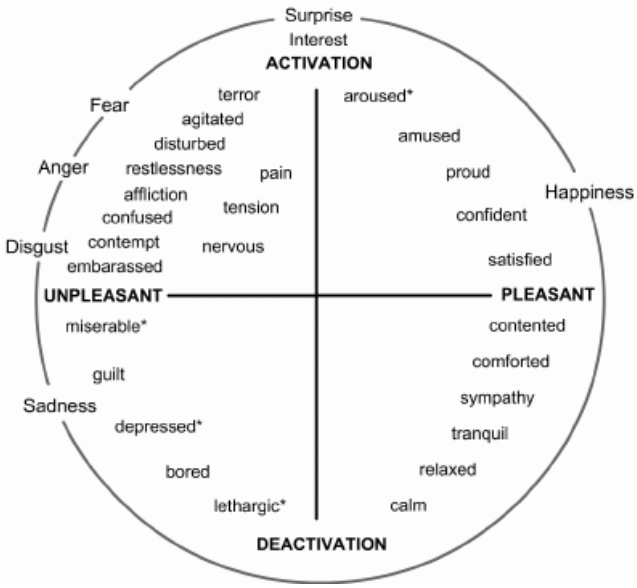

Figure 3: Circumplex model of Emotion (Zagalo et al., 2005). The horizontal plane represents valence, and the vertical plane represents arousal/activity.

Daly et al. (1983) third dimension of intensity, the two are differentiated, with contempt being coded as high intensity, while agitation is represented by low intensity (Guerrero et al., 1996). Yet further dimensions can be added to better "fine-tune" the model, as show by Plutchik (1980) multidimensional model, which addressed basic emotions and how they might be arranged into hierarchies. To do so, he added polarity (in which eight primary emotions are arranged in opposing pairs across from each other), similarity (where adjacent pairs reflected emotional similarity), and intensity (represented by the vertical levels of the model, such as joy varying from ecstasy to pleasure) (Guerrero et al., 1996). By adding these dimensions, Plutchik (1980) allowed for the complex relationships between emotions to be observed and analyzed. With the addition of more dimensions, emotional synthesis can be more readily fine-tuned to work off the myriad of attributes indicative of emotional change. However, emotional recognition benefits inversely from added dimension. While an increase of dimensions may allow for more accurate classifications, it may lead to diminishing returns, where too many dimensions may increase the risk of problems occurring due to the complexity of the decision-making software. Perhaps because of this issue, most models tend to work with two to three dimensions to avoid such a risk.

Although both approaches have their benefits and shortcomings, they are not necessarily exclusive to one another. Researchers have agreed that to best broach the idea of emotional research, a method of using both approaches to cover the other's faults is best to accomplish wide-reaching studies (Harmon-Jones et al., 2017). One study focused upon utilizing automatic nervous system (ANS) patterning to differentiate discrete emotions, such as joy, anger, fear, and sadness, and group them via a 2-dimensional circumplex model of valence and activation. In doing so, they found that the hybrid model based upon the ANS patterning allowed for self-reported emotions to be accurately located within the circumplex model and connect the ANS activity also found in the model to specific emotions (Christie & Friedman, 2004). This indicates further discriminatory applications of using a multivariate approach of emotional classification. Another such study attempted something similar,

this time mapping discrete emotions into a three-dimensional mode based off pleasure (valence), arousal (activation), and dominance (emotional intensity) (PAD) (Hoffmann et al., 2012). While they did not find that there was a possibility of a universal mapping between discrete emotions and the PAD model, they realized that a more flexible transition between the two classifications based on individualized calibration would benefit Affective Computing research more. Despite the setback to development of a universal transition software between the two approaches, individual assessment is more reliable in obtaining such an idea. Smaller scale studies, such as this one, may in fact be a component in aiding drawing that bridge between the two approaches. However, the two approaches are still, for the most part, conceptual and subjective in nature. Speech and emotion retain analytical attributes, that are more easily utilized in emotional recognition and synthesis, of which have been briefly touched upon previously.

Previously, we have been looking at the conceptual approach to emotion classification, or what might also be described as perception-oriented studies, those that are concerned the ability of listeners to identify emotions. However, the more analytic approach, acoustic-oriented studies, which are concerned with the analysis of acoustic features, is what a majority of emotion recognition and synthesis models use to best achieve their goals (Razak, Abidin, & Komiya, 2003). Speech features tend to be divided into two main categories: phonetic features, such as vowels and consonants, and prosodic features, such as rising and falling tones and stress. For the most part, emotion-based studies depend more on the prosodic features, such as pitch (fundamental frequency), and amplitude (loudness). Multiple forms of research have been done over the years in order to determine what values best capture each emotion. Generally, the emotions looked at, tend to be those described as the "basic" emotions explained earlier. These include fear, joy, sadness, and anger, but are not limited to these four. For instance, Sobin and Alpert (1999) affirmed that fear was characterized by increased pitch, pitch variance, and speaking rate, while sadness was manifested by lowered pitch, reduced volume, and slowed speaking rate. Anger was characterized by higher volume, volume variance, increased pitch variance, and rapid speech rate, but contrastingly to other studies, displayed low pitch rather than high. Furthermore, joy was similar in anger, in higher volume, pitch variance, and speech rate, but low pitch. This difference in the norm, they stated, might have been due to an all-female sample for their studies. This, however, contrasts Kolita and Acharjee (2021) study on emotional features in the Assamese language, where they found that woman had larger mean pitch values in angry emotions, at about 236Hz-475Hz, than in normal/neutral emotions at 186-285Hz. Yet, another such study found that anger, for instance, had an average pitch in English at about 160Hz for females, but did not differ much from fear, which runs counter to Sobin and Alpert (1999) findings (Razak et al., 2003). This could be explained by language specific reasons, such as higher average pitch found in the Assamese language or by improvements or differences in technology. While these papers show vastly different results for attributing values to speech features over time, these features have been generalized, or at least accurately extracted and recorded to a point that a mass amount of emotion representation models depend on these values to either recognize or synthesize emotion consistently. Yet, despite the extensive use of speech features and the acoustic-oriented approach, simple and generalized models still work effectively while using a more perception-oriented approach, which may find more use when dealing with emotion found in text, rather than with speech. The best course of action, then, is to work with multiple models, each running with a particular approach, in order to cover for the other approach's shortcomings, as will be evident in the following sections.

2.2 Emotion Recognition

Emotional recognition, the process of automatically identifying and categorizing emotional states from speech and text data, stands at the forefront of affective computing research. In this section, we delve into the multitude of models and approaches to answering the problems that underlay the field and impact human-machine interactions.

Within the realm of emotional recognition, models incorporating emotional cues and feature extraction mechanisms play a central role in deciphering the subtle nuances of human affective expression. As mentioned in the previous section, speech contains variable cues and features that are instrumental in portraying and determining emotion. As such, it is clear that to best predict and recognize a particular emotion, a machine model would need to analyze and process such features to fulfill its task. While each feature is important to fully recognize emotions in speech, researchers have tended to focus on only a few at a time, increasing the efficiency of the model's recognition function. The most widely used features to best predict and classify emotions seem to be mel-frequency cepstral coefficients (MFCCs) and pitch, with loudness finding more use in anger-based emotions (Gunes, Schuller, Pantic, & Cowie, 2011).

A., V.R., A., Jayakumar, and P. (2009) outlines this well, explaining the efficiency that MFCCs have in representing both the linear and non-linear properties of speech. Their study continues, turning to Discrete Wavelet Transform coefficients (DWT), incorporating multi resolution filter banks for the analysis of signals. After using a feed forward neural network on a dataset of four emotional classes (neutral, happy, sad, anger), A. et al. (2009) found dividing results, where MFCCs recognized happiness and anger at an accuracy of 80% and 95% respectfully, while DWT preferred neutral and sad emotions at 68% and 60% respectfully and achieving a 92% accuracy for anger. However, the two features struggled in the emotions they did not exceed in, resulting in an overall accuracy of 68.5% for DWT and 55% for MFCCs. While these results might indicate that these features were not as important or as required as many other researchers believed, their individual emotion results still show that there is a point of progress and further research.

This continues further, as Polzehl, Schmitt, Metze, and Wagner (2011) decides to focus on a particular emotion (anger) and determining the best combination for features and recognition. Their process begins by applying large-scale feature extraction, capturing the expression via calculation of prosodic and acoustic features. After deriving statistics from the features, they fuse information from both linguistic and acoustic classification results to obtain an estimate of the emotional state given. Running three databases through this process (2 German, 1 English), they found that the interactive voice response datasets (IVR; English and German) returned good results, while the German Wizard of Oz (WoZ) dataset seemed problematic, which was believed to be due to mapping multiple classes into a single cover class of Anger, causing blurring in the acoustic models. Nevertheless, the results displayed that MFCCs and loudness seemed to account for more than 50% of all the features, while pitch accounted for approximately 25% for the English IVR database. After fusion, the WoZ database returns a 75.3% accuracy and a .70 f1, and the German and English IVR return a 78.9% and 78.2% accuracy with a .78 f1 between the two. This indicates further assurance that features chosen, such as MFCC and pitch, do in fact hold onto that semblance of usefulness. Furthermore, the combination of the acoustic model and linguistic model, which was lexically based, resulted in a better accuracy, rather than the linguistic model on its own.

This focus on linguistic features retains in other languages, as was displayed in Polzehl et al. (2011) study, utilizing and recognizing emotion in German datasets. Stepping away from the focus

on MFCCs and large-scale extraction, Kolita and Acharjee (2021) and Mohanta and Mittal (2017) used linguistic features as support for their overall studies in the Assamese and Telugu languages respectfully. As mentioned in the previous section, Kolita and Acharjee (2021) used the comparison and analysis of pitch and format as a way to evaluate the differences in emotional effect with reference to biological sex. Utilizing three emotional styles, -anger, normal, and surprise- the mean pitch and format frequency were calculated and analyzed amongst male and female voices in the Assamese language. After testing the results via a Two-way ANOVA, Kolita and Acharjee (2021) determined that Angry emotions resulted in higher pitch and greater formant values compared to normal emotions, and that female speakers on average sustained higher mean values compared to male speakers. While the results are not groundbreaking in nature, it solidifies the variability and interconnected nature of emotions and linguistic features. Mohanta and Mittal (2017) attempted to classify four emotions, happy, anger, fear, and neutral, by analyzing changes in features in the vowel regions. Interestingly, they bring up that "apart from the handful number of basic emotional states [...] there are many other complex emotional states, which are not possible to identify using only the speech signal" (2017). They continue by stating the two methods of human recognition, via physical expression, such as facial, body posture, etc., and human speech signal, or more specifically, the meaningful pattern found in the signal. While Mohanta and Mittal (2017) extract pitch and formant features, their focus is on classification emotions with only the vowel part of the speech signal. The values of pitch and features give measurable and variable properties to the vowels chosen, so that the classification can refer to such values when training and making its prediction. After training their data on Support Vector Machine (SVM) classifier, the accuracy of classification resulted in 76% for happy, 70% for anger, 60% for fear, and 75% for neutral emotions. While this baseline was achieved only through a combination of vowels and their subsequent feature values, this study shows the variability in emotional recognition.

While speech is the primary sample in emotion recognition models, as mentioned above, physical expressions can do the same, if not more. Gunes et al. (2011) state that in reference to arousal and valence dimensions, physical features such as facial expressions and body postures form widely known and used visual signals for automatic affect analysis (recognition) and synthesis. With the use of motion capture technology, mapping facial expressions and movement can be trained to predict emotional states. Busso et al. (2008) does just this, in which they capture both facial expressions and speech data from multiple actors to best recognize emotional states, and compiled and classified it into a database. In order to do this, there was a focus on naturalness, so that the emotions could be naturally elicited, and emotion labels were determined by human subjective evaluations. The data was divided and categorized into six basic emotions: happiness, anger, sadness, neutral, disgust, fear, excitement, and surprise, mirroring closely to Ekman and Cordaro (2011) definition of basic emotions. From there, evaluators also agreed upon scores for valence, activation (arousal), and dominance, the definitions described in the previous section. This database, while not a model, is extremely rich in its multi-faceted approach in emotion classification and recognition, and was created, in the words of Busso et al. (2008), to "[...] play an important role in understanding and modeling the relation between different communicative channels used during expressive human communication and contribute to the development of better-machine interfaces". It is no wonder, then, that this database is used in training both emotion recognition and emotional synthesis models, such as the one to be used for this study. However, physical and sonic aspects are not the only source for emotional recognition. How do we handle sentiment and emotions found in text, such as fictional novels?

Text-based emotion recognition differ from other emotional recognition models in that they require semantic and sentimental data to properly predict. The models we have been previously reviewing work by extracting linguistic features, however text-based models don't have the benefit of working with audio. To solve this issue, three approaches were found to allow machines to recognize emotions, or sentiment, in text: keyword recognition, lexical affinity (LA), and machine learning (ML) (Acheampong, Wenyu, & Nunoo-Mensah, 2020).

Keyword recognition deals with the construction and use of emotion dictionaries or lexicons (Acheampong et al., 2020). The emotion lexicons contain sentiment words such as happy, angry, excited, etc., where once the keyword is identified in a sentence, a label -usually the emotion- is assigned to the sentence. How these keyword recognition models are trained follow two roads of thought: Hidden Markov Models (HMM) and Dynamic Time Warping (DTW). The HMM model uses statistical analysis to obtain acoustic representations of keywords from training data while DTW find optimal warping path between the start and end points of a text or speech segment and the representative keyword utterance (Park, Jang, & Kim, 2012).Though it is simple, issues arise from this approach, such as the requirement of an emotion dictionary to contain a reasonable number of categories. Nevertheless, researchers have attempted to overcome these shortcomings, determining novel ways to enhance keyword recognition, or as described in many studies, keyword spotting.

Park et al. (2012) tackles one of the issues of keyword recognition, focusing on an utterance verification module, which decides if a candidate keyword segment should be classified as such. Proposing a keyword verification technique characterized by multistage utterance verification, Park et al. (2012) surmised that removing log-likelihood results for garbage models (non-keyword results) from the ranking of all models determined in an HMM-based recognition system would reduce the number of errors the model would make. An alternative to this was to evaluate the distance between the log-likelihood of the first rank and the last rank of the recognition results, ignoring all the middle ranks and the garbage models. Afterwards, the results would be pushed through a DTW algorithm as a second verification stage. After training and comparing four HMM models, -two conventional and two of their proposed- they found that their proposed models improved the accuracy of detecting keywords and reducing the error rate by about 8%. This study, while to minor effect, overcomes one of the issues of keyword recognition, dealing with false positives and false failures.

Another such study handles the prevalent issue of Out of Vocabulary (OOV) words with a different form of a multi-step verification process. Santoro and Marcelli (2019) determined that a major issue with transcribing historical handwritten documents was that the keywords relevant to the document were obtained by transcribing documents of the training set and thus not representing the complete list of keywords of the documents. The proposition as follows was a human-in-the-loop method, meaning that as the keyword dictionary is being created, the user of the model will be shown the keyword with its transcription during the validation step. Following that, the user can decide to validate the correct outputs, correct the wrong outputs, or transcribe manually the OOV words. After testing the method, the results confirmed that by transcribing the OOV keywords along the process, leads to a significant reduction of time required to transcribe the document. Furthermore, they found that though using a keyword spotting system improved performance in terms of time gain, the improvement was bounded by the size of keyword list. This study solved one of the more glaring issues of keyword recognition, albeit in a unique way. However, the requirement for human interference still lowers the evaluation of keyword recognition models.

The second approach, lexical affinity (LA), serves to augment the keyword recognition method. LA assigns probabilistic affinities, such as "positive" or "negative", supplementing the identification

of emotions via keywords (Acheampong et al., 2020). However, this forces the emotions into two extreme states, shown earlier. This leads to inaccuracies in classification depending on the context of the keywords. While not as used widely, some models touch upon LA in order to simplify emotional recognition for larger usages.

The model explored earlier, by Polzehl et al. (2011), uses the LA approach to some degree of success, as they are focusing only on a single emotion: anger. After compiling the categorizations into anger and non-anger, the level of the anger connotation of a word can be subsequently estimated, without the risk of further inaccuracies via LA. While the study retained decent and comprehensive results, they were much lower and closer to the baseline of 60% overall accuracy when only working with the linguistic model. The added need for an acoustic model to improve the results indicate that issues with LA based recognitions are still prevalent.

The final approach, machine learning (ML), utilizes algorithms to aid in emotion category classifications. While supervised ML algorithms have been widely implemented for text-based emotion recognition problems and result in generally better detection rates than unsupervised models, unsupervised techniques have slowly begun to see more use, such as the support vector machine mentioned in Mohanta and Mittal (2017) (Acheampong et al., 2020). However, unsupervised methods are not as robust to effectively detect emotions in texts, in the end paving the way for supervised deep learning models, as their deep lays can extract hidden details that text may carry. As research continues, hybrid approaches have begun to come out of the woodwork, combining rule-construction and the ML approaches into a singular model. This leaves much to look forward to, as the hybridity can begin to compensate for the issues each part lacks.

Bharti et al. (2022) broached this idea of hybrid models, mixing both ML, in the form ML classifiers, and deep learning, in the form of deep learning models. Their proposition was that their hybrid model could detect emotions that are "tasteless", or do not have any tone or expression. By sending the pre-processed data through multiple models, choosing the best of those resulting models, then combining them into a final, evaluated hybrid model, Bharti et al. (2022) end up choosing a SVM as the ML classifier and a convolutional neural network (CNN) and Bidirectional Gated Recurrent Unit (Bi-GRU) as the final pieces of the completed model. On their own, the SVM returned an accuracy of 78.97%, the CNN an accuracy of 79.32%, and the Bi-GRU a 79.42%. When combined, the hybrid model resulted in improvements across Precision, Recall, F1 score, and accuracy, with it ending at 80.11%, showing an evident, though minor, improvement in the model's capabilities, indicated the usefulness of hybrid models.

Another such model, and the basis of the model used in this study, is the Bidirectional Encoder Representations from Transformers model, or BERT (Devlin, Chang, Lee, Google, & Language, 2019). BERT is a multi-layer bidirectional language representation model, pre-trained on two unsupervised tasks: Masked LM, i.e. some percentages of the input tokens are randomly masked/hidden, and Next Sentence Prediction, which is explanatory in the name. Fine-tuning was then done, focusing on text pairs, taking the forms of pairs in paraphrasing, hypothesis-premise pairs, question-passage/answer pairs, and a degenerate text pair for text classification. Compared to other models without bidirectional MLM, BERT outperforms in predictive accuracy. However, Y. Liu et al. (2019) believed that BERT was undertrained and not fully representative of its capabilities. To rectify this, they proposed four modifications: 1) training the model longer, with bigger batches, over more data; 2) removing the next sentence prediction objective; 3) training on longer sequences; and 4) dynamically changing the masking pattern applied to the training data (Y. Liu et al., 2019). Because of the ease of which BERT can be fine-tuned, via inputs of two segments, or sequences of tokens, pre-

training and subsequent tuning was done with ease over five English-language corpora, averaging at about 160GB of uncompressed text. Robustly optimized BERT approach, or RoBERTa, resulted in a large improvement over the original BERT results, and exceeds other systems, such as XLNet, on multiple leaderboards.

RoBERTa eventually became the basis of the emotion recognition model used for this thesis: Emotion English DistilRoBERTa-base (Hartmann, 2022), or EED. The EED model is a distilled version of the RoBERTa model, focusing on emotion label pairs, trained on a diverse collection of text, including texts from Twitter, Reddit, utterances from TV dialogues and Multimodal EmotionLines Dataset (MELD). The model resulted in a 66% evaluation accuracy, and subjectively evaluated by a human at an average of 75% over all emotional categories in its task to predict Ekman and Cordaro (2011) 6 base emotions plus a neutral class. The straightforward and ease of use and implementation of the model set it apart from others, as it can be used to iterate through large pieces of text quickly and efficiently to later be used in other forms of code, such as emotional synthesis.

2.3 Emotion Synthesis

Speech Synthesis, or generally known as Text-to-Speech (TTS), has been a common point of research for human-machine interaction over the years. To make conversation more natural, machine learning models consistently show up each year in an effort to make synthesized speech and conversation more natural. Above all, this includes expressive and emotional speech, to allow the synthesized voice to become that much more "human" or pleasing to human ears (Shen et al., 2018). However, before diving into the models that have found ways to properly synthesize emotional speech, two TTS models that serve as a basis for other models to compare to or be built upon need to be explored: Tacotron2 and FastSpeech2.

Tacotron2 is a neural network architecture, built to improve upon the issues of the original Tacotron, by generating mel spectrograms, followed by a modified WaveNet vocoder (Shen et al., 2018). The two parts of Tacotron2 compliment each other nicely, covering for the other's shortcomings. For instance, while WaveNet consistently produces audio quality that can rival real human speech, it requires significant domain experience to produce, including elaborate text-analysis systems and robust lexicons. Tacotron, a sequence-to-sequence architecture, simplifies traditional speech synthesis pipelines by replacing the production of speech (linguistic and acoustic) features with a single neural network. However, it depends on the Griffin-Lim algorithm for vocoding use, which characteristically produces artifacts and low audio quality (Shen et al., 2018). As shown, the issues from each model compliment one another nicely. To link these two components, mel-frequency spectrograms (MFS) are used. Using the MFSs allows for details in lower frequencies to be emphasized, while details in higher frequencies are de-emphasized, reducing the amount of noise bursts, but increasing intelligibility. After training for 50,000 iterations and evaluating on 100 fixed samples, the model was found to generate high-quality audio using as few as 12 layers, compared to the 30 layers of the baseline model (Shen et al., 2018). However, the model still suffers from slower inference speed and work skipping/repeating issues.

To combat these issues, Ren et al. (2022) created FastSpeech2, a non-autoregressive model which focused on speed and voice quality. To overcome the obstacles, FastSpeech2 incorporated three changes from its original model: [1] impler training pipeline, which results in faster training time, [2] achieving better voice quality by alleviating the one-to-many mapping issue, [3] simplify inference pipeline, while also maintaining high voice quality, by directly generating speech waveform from

text. In order to simplify the training pipeline, Ren et al. (2022) directly trained the model with the ground-truth target instead of the simplified output of the teacher, as the two-stage teacher-student process made training complicated. Furthermore, they introduced variation information of speech, such as pitch and energy, and took them as conditional inputs to reduce the information gap that occurred along with the mapping problem characteristic of non-autoregressive TTS model training. Comparing it to other models, such as Transformer TTS and Tacotron2, FastSpeech2 saw similar mean opinion score results to conventional TTS systems, but also improved training time by almost twice compared to the Transformer TTS and more than three times compared to the original FastSpeech model Ren et al. (2022). Furthermore, an additional feature of FastSpeech2 that allows further prosodic customization is variable pitch, energy, and duration control, allowing for like-expressive results in the synthesized audio.

Due to the relative recency of the creation of FastSpeech2, most studies on expressive speech in the past several years have opted to use Tacotron2 as their base architecture, to positive results. Lei, Yang, Wang, and Xie (2022) proposed a multi-scale emotion speech synthesis model with three proposed modules, based on style transfer TTS, in which the aim is to synthesize speech with the same style as that of the reference audio. Their three modules were as follows: [1] global-level emotion (GM) which is concerned with what emotion category is conveyed by the entire utterance, [2] utterance-level emotion (UM) which focuses on the prosody pattern within an utterance. [3] local-level emotion (LM) which provides emotion strength for speech pronunciation (syllables and phonemes), can control the intensity of localized emotional expressions. Using the encoder part of Tacotron and the decoder part of Tacotron2, and trained on a corpus which contains six kinds of emotional speech for about 12 hours of audio, Lei et al. (2022) found that their model resulted in a lower mel-cepstral distortion (the difference between synthesized results and ground-truth speech) than the GST model, a state-of-the-art reference audio embedding method. Overall, their model, called MsEmoTTS, achieves good performance via transferring, while also allowing for the option to be controlled manually to synthesize emotional speech at GM and LM levels.

Another study that shows the controllability of Tacotron2's architecture is Luo, Takamichi, Saito, Koriyama, and Saruwatari (2024) emotion-controllable synthesis using soft labels and prosody factors. They proposed a paired speech emotion recognizer (SER) prosody factor generator (PFG) that estimates emotion soft labels from utterance-level prosody factors, and vice versa, that can produce a combined coarse-grained (emotion labels; happy $\rightarrow$ angry) and fine-grained (emotion strength; pitch and energy mean, and range) control of speech emotion. Although their results showed that their method regarding emotion-perceptual accuracy was inferior to conventional approaches that cannot control the emotion of the synthetic speech, the model nonetheless performed at the same level as those approaches, despite only being trained on a weak emotional dataset. While the results are not stellar in any means, the promise that this study shows may allow for further research to tend towards completely controllable expressive speech synthesis.

This fine-tuning of emotional synthesis continues to what seems to be a new step, in emotional mixing in speech synthesis. Zhou, Sisman, Rana, Schuller, and Li (2023) uses Plutchik and Kellerman (1980) 8 primary emotions and emotion wheel to guide their model's output. To do this, they study the relative difference between emotional categories, which can be labeled as an emotional attribute, then adopting a text-to-speech framework with joint training of voice conversion. Their point of research followed mixing Surprise with three other emotions, Happy, Angry, and Sad, which was expected to synthesize Delight, Outrage, and Disappointment, which they believed was easiest of listeners to perceive. After training, subjective evaluation shows that the model succeeded in syn-

thesizing new nuanced emotion types that are more difficult to find in real life. While the study only focused on emotional TTS, the results provide many potential improvements to current emotional speech synthesis frameworks, bringing us closer to emotional intelligence. This study is pinnacle in showing a more perception-oriented base for emotional synthesis, focusing on the discrete approach mentioned in the previous section. However, despite allowing for emotional mixing to occur consistently, the same issues that plague the discrete approach can be transferred to this study. While adding or removing more of a particular emotion to the mixture can allow for reproduction of the relevant emotions on the wheel, the results are still much too broad and does not allow for more minute changes that are representative in dimensional approaches.

In parallel, Y. Lee, Rabiee, and Lee (2017) presents an end-to-end neural speech synthesizer designed to generate emotionally expressive speech using the Tacotron model. They address key challenges in the original Tacotron, such as exposure bias and irregular attention alignment, by integrating context vectors and implementing residual connections in the recurrent neural networks. Their enhanced model, known as emotional Tacotron, effectively incorporates emotional embeddings, allowing for synthesis that accurately reflects specified emotional states. Results demonstrate that these modifications improve the clarity and sharpness of attention alignments, leading to higher quality and more intelligible speech outputs.

Building upon this concept, the same approach is adapted to modify the FastSpeech2 model, resulting in an Emotional TTS system capable of conditioning both continuous emotional descriptors (such as arousal and valence) and categorical descriptors (such as basic emotions like happiness and anger) (K. Lee, 2021). This adaptation allows for a more nuanced and flexible emotional speech synthesis, leveraging the robustness and efficiency of the FastSpeech2 architecture while integrating sophisticated emotional controls. By conditioning on continuous descriptors, the modified model can produce speech that varies subtly across a spectrum of emotional intensities, providing a more nuanced output. Simultaneously, the use of categorical descriptors ensures that the synthesized speech can distinctly convey primary emotions, enhancing its use in providing more specific and complex renditions of emotional speech. This dual conditioning capability draws in parallel with a hybrid model that works with both the strengths of discrete and dimensional approach of emotion classification, in turn providing advancement in emotional speech synthesis.

3 Methodology and Experimentation

Creating a generalized process for transitioning emotional text to accurate and reliable emotional speech requires a system in place to efficiently execute its function. In this study, two datasets were used for training and testing the two chosen models: Emotion English DistillRoBERTa-base and the Expressive-FastSpeech2 model. The lines were then categorized and processed into intelligible audio. Finally, twenty human participants were asked to evaluate through a survey the emotion and intensity of select audio samples, of which eight were synthesized as neutral and twelve were synthesized as anger. Each of these steps, as well as the ethical considerations of this study, will be outlined in detail in the sections below.

3.1 Databases

Two datasets were used to compile this experiment and automation. The object of study and experimentation came in the form of Alice in Wonderland by Carroll (2006). The book was found on project Gutenberg (an online library of over 70,000 eBooks) and was edited so that the preface, acknowledgments, and license sections were removed from the text. Any placeholders for illustrations were removed as well. Once only the text of the novel remained, the novel was split by new lines, allowing for subsequent models to iterate over them. An idea was attempted to split by dialogue/quotations, however as future sections will explore, the resulting emotional classification returned poor results.

The synthesis model was trained with the IEMOCAP, or "interactive emotional dyadic motion capture", database, which is an acted, multimodal and multispeaker corpus, collected and compiled at SAIL lab at USC (Busso et al., 2008). It contains 12 hours of audiovisual data representing both improvised and scripted sessions, recorded via five female and five male actors, chosen to display emotional expressions in both audio and video formats. The dataset is labeled and annotated into emotional categories by six human evaluators who majoritively agreed on all the categorizations, which includes happiness, anger, sadness, frustration, disgust, fear, excitement, surprise, and a neutral state, but is also labeled in valence, activation, and dominance, displaying a wide range in how the different categorical emotions are elicited. The database was pre-processed for use via Montreal Forced Aligner (MFA) at a sampling rate of 22050HZ and 80 mel-channels with a mel-frequency maximum value of 8000. MFA was first trained with the database from scratch, after which the database was aligned once again with the trained MFA and extracted lexicon dictionary.

3.2 Models

Two models were used in this automation. For emotional classification via text, the Emotion English DistilRoBERTa-base[1] (EE-Distil) fine-tuned by Hartmann (2022) was used, as it follows the same 6 basic emotions that were considered via Ekman's theorem, plus an extra neutral class that aided in further classification. The model derives from a distilled version of the RoBERTa-base model, a transformers model pretrained in a self-supervised manner on a large corpus of English data in order to be used for masked language modeling. The distilled version contains 6 layers, 768 dimensions and 12 heads, for a total of 82 million parameters, but works on average twice as fast as the base

[1] https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/

version despite having less parameters. Hartmann's fine-tuned model was trained on a balanced subset of near 3000 observations per emotion, derived from emotion labels from Twitter, Reddit, student self-reports, and utterances from TV dialogues. Another model that was fine-tuned further on transcripts from the Friends TV show, trained by Li (2023)[2] was also considered, however, in human-based comparative subjective tests between the two models done on Alice in Wonderland, Hartmann's model displayed an average correct classification of 75.47% compared to Li's model's 72.60%. While percentage differences were minimal, internally, Hartmann's model lacked in determining texts that reflected the emotion Joy, compared to Li's model, though did well in determining "negative' emotions such as Anger, Disgust, and Fear, of which Anger was the focus for this study.

The model used for synthesis is K. Lee (2021) Expressive-FastSpeech2[3], an emotional TTS model based off of FastSpeech2 as a non-autoregressive multi-speaker TTS framework. FastSpeech2 operates by first predicting the duration of each phoneme in the input text and then generating mel spectrograms based on these durations. The mel spectrograms are subsequently converted into waveforms using a vocoder such as HiFi-GAN, which was the default vocoder of the Expressive-FastSpeech2 model. FastSpeech2 introduces a variance adapter to control prosody elements such as pitch, energy, and duration, allowing for more expressive and natural speech synthesis. This model forms the foundation of our synthesis approach by providing high-quality, fast, and flexible speech generation. Expressive-FastSpeech2 expands upon this by following implementations found in the Emotional End-to-End Neural Speech synthesizer study. These implementations focus on emotion embedding that is conditioned in the utterance level, after which, emotion, arousal, and valence, based on the dataset, are employed for the embedding. The concatenated embedding passes through a single linear layer with ReLU activation for the fusion, which is then processed by the decoder to synthesize speech in the given emotional conditions (K. Lee, 2021). Using similar parameters from the original FastSpeech2 model, the Expressive model's transformer contains 4 encoder layers and 6 decoder layers, both with a hidden size of 256 and a dropout rate of 0.2. After, a convolutional layer with a filter size of 1024 and kernel size of [9, 1], followed by variance predictors with a filter of 256, kernel size of 3, and dropout rate of 0.5 is applied. The new parameters that separate the original model with this one is the Boolean value for multi-emotion is set as 'True' and an increased maximum input sequence length of 2000.

3.3 Audio Processing

The Alice in Wonderland text was split by new lines, allowing for large portions of the text to be iterated through. The lines of text contained both expositional context, dialogue, and emotion adjectives that labeled the dialogue as the given emotion. An alternate idea was tested where the text was split by quotations, thus keeping only the dialogue in its entirety. However, results from iterating the EE-Distil model showed confusing and inaccurate classifications, so the entirety of the line was kept as is. The entirety of the text was iterated through by the model, in which each line was classified and stored in a separate text file based on the highest rated emotion and compiled with their scores for their top three emotions for later use.

The Expressive-FastSpeech2 model was trained for 900,000 steps over the course of 5 days, saving every 10,000 steps for future analysis and refinement of checkpoints. Learning rate was

[2]https://huggingface.co/michellejieli/emotion_text_classifier

[3]https://github.com/keonlee9420/Expressive-FastSpeech2

reduced via anneal steps at 300,000, 400,000, and 500,000 at a rate of 0.3. Validation was done every 1,000 steps in order to monitor overfitting and generalization more frequently. After analysis of saved checkpoints in validation logs, step 830,000 was chosen as the model to be represented by. Pilot tests allowed for further refinement in choosing a particular speaker from the IEMOCAP dataset to conduct all our experiments with, giving consistency to the synthesized output. One speaker was chosen, a Female speaker from the third session of the IEMOCAP database.

Further pilot testing and experimenting with the effect of changing the valence (negativity or positivity of an emotion) and arousal (level of excitement in an emotion) took place. Both valence and arousal attributes were scaled from 1 to 5 [valence: 1-negative, 5-positive; activation/arousal: 1-calm, 5-excited] (Busso et al., 2008), increasing by either quarter steps (.25) or by thirds (.3333), allowing for minute adjustments and refinement to the output emotion.

After the effects of the changes in arousal and valence were understood, two text files, categorized by emotion, were chosen: anger and neutral. The anger-coded lines were further cut down to only include lines that scored at least a 50% on their anger probability, and then organized in order from highest to lowest. From there, the scores and their connected lines were sectioned off into three categories defined by intensity: 50-65% anger scores were categorized as low (intensity) anger, 66-81% were moderate anger, and 82%+ was categorized as high anger. The three sections best represented the varying scales of anger, as displayed on Zagalo et al. (2005) circumplex model of emotion. Afterwards, 4 lines were chosen from each section of the anger text, to allow for diversity in the content of the text, and 8 lines from the neutral text were chosen at random, allowing for a "standard" for the anger-coded lines to be compared to. The chosen lines can be found in Appendix A. Each chosen line was then reduced further to synthesize the dialogue portions of the lines, removing any exposition or descriptors of what the line was labeled as. After short experimentation and judgment from the main researcher and a critical friend with experience in teaching emotional expression in theater and music, values for the valence and arousal attributes were chosen for each "category". As neutral was given a score of 3.0000 for both valence and arousal, the resulting anger category had to work within a threshold of 1.0000-3.0000 for valence, and 3.0000-5.0000 for arousal. The values found for valence and arousal for each section were as follows: Low anger: Arousal=3.5000, Valence=2.3333; Moderate anger: Arousal=4.3333, Valence=2.0000; High anger: Arousal=5.0000, Valence=1.0000. While scores seem close together, such as low anger and moderate anger's valence scores, the resulting speech was diverse enough to tell the difference between the two categories.

Due to the presence of noise in the resulting audio samples, the synthesized files were put through two noise/voice fixers, then rejudged by the researcher and critical friend to determine the most usable sample: VoiceFixer (H. Liu et al., 2021) and Adobe Podcast Enhance (Adobe, 2022). While most of the audio samples were cleaned-up via Adobe Enhance, several samples benefited more from VoiceFixer instead.

3.4 Survey Analysis

A survey was created to test the capabilities of the model synthesizing nuanced anger neutral speech. Twelve anger and eight neutral audio samples were compiled and randomized in position. Participants were asked to listen to the audio files and answer questions on them one at a time. The questions asked the participant to determine what the primary emotion in the audio sample was, and at what intensity the emotion was at. Intensity was defined on a 7-point scale, with 1 being low in-

tensity and 7 being high intensity. A 7-point scale was chosen to minimize the amount of assumption required to analyze the participant's choices. With 7-points, the positions of 1 – low, 4 – moderate, and 7 – high, allowed for two transitionary choices between the extremes and the moderate positions. The two transitionary choices filled the roll of "leaning" choices, aiding in the interpretation of what the participant's intentions were when making the choice. While intensity was explained earlier as a distinguishable attribute to valence and arousal, for the sake of ease of understanding for non-academics, intensity was used as a generic term to rate the overall differences of each stage of emotion. Furthermore, participants were given the option to choose an additional emotion if they believed there to be multiple emotions or were stuck between multiple possible emotions. This allowed for the participant to still have a chance in correctly recognizing the intended emotion, in case they did not choose anger or neutral as the primary emotion for their current audio sample.

A small sample of participants was chosen to be part of the pilot survey, to aid in finding faults in the survey as well as allow the researcher to determine the best method for data analysis. After the pilot test, the survey was distributed across several platforms, such as Discord and Reddit, as well as acquaintances of mutual connections and made available for a week to complete. After the week had passed, a sample size of 20 was achieved. An example of the survey questions can be found in Appendix B.

3.5 Ethical considerations

As the database, IEMOCAP, contains the voices and faces of human actors participating in the study that compiled the data, and the evaluation process of this thesis is subjective in nature, and required human participation, several matters of ethical implication must be taken into account.

In compliance with the provider of the IEMOCAP database, no pre-trained models will be shared to facilitate consistency or ease of implementation across the field, and once the experiment is finished, all forms of the database are deleted from the programs it was used with, as well as any storage devices it was held in, in order to avoid the risk of possible data breaches or theft.

In the case of the subjective evaluation, Qualtrics was used to create and compile the survey and its results. The survey was made in a way that no information by the participants would be recorded, save for their level of English. The survey was spread with an anonymous link, that recorded no information, including IP addresses, to further solidify the anonymity of the study responses, and to reduce the risk of their personal information being leaked in the case of a data breach. A copy of the consent form can be found in Appendix C

Additionally, despite the survey being distributed to personal or mutual connection, there is no way of knowing who participated in the survey, unless the participant themselves mentioned it, and there is little risk of bias in analyzing the results, as there is no way to determine whose responses were which.

Replication of the research can be done by following the links to the GitHub repository and Hugging Face model, however the IEMOCAP database must be formally requested to be used by the providers.

In the next section, the results of the survey and their impact on the study will be elaborated.

4 Results and Discussion

In this study we have tested to see how well participants could recognize the emotion labeled in the audio and the subsequent "intensity" level of the labeled emotion, which in this case was Anger and a Neutral state. Participants were given two chances to correctly deduce the correct emotion and allowed 'transitionary' scores as explained earlier in the previous section. Analysis was split up into different portions, dividing results between emotion recognition and intensity recognition, which then was further divided into first and second tries.

4.1 Emotion Recognition Analysis

Percentile-based analysis was done to determine overall recognition accuracy for the basic emotion and the intensity level of the audio sample. Across 20 samples, we see an overall average of 48.75% of participants accurately recognizing either Anger or Neutral in a given audio sample. Divided further by emotion, 40.83% of participants were able to correctly recognize Anger across all audios, while 60.63% were able to correctly recognize the Neutral state. These percentages reflect responses where the correct emotion was chosen the first time, in reference to the study, the "primary emotion". When incorporating the second try, or "secondary emotion", we see an increase in percentages. The overall average became 55.25%, while Anger-specific recognition became 47.92% and Neutral became 66.25%, indicating a better than random chance result for recognizing the emotion in an audio if given the opportunity to label it with multiple emotions.

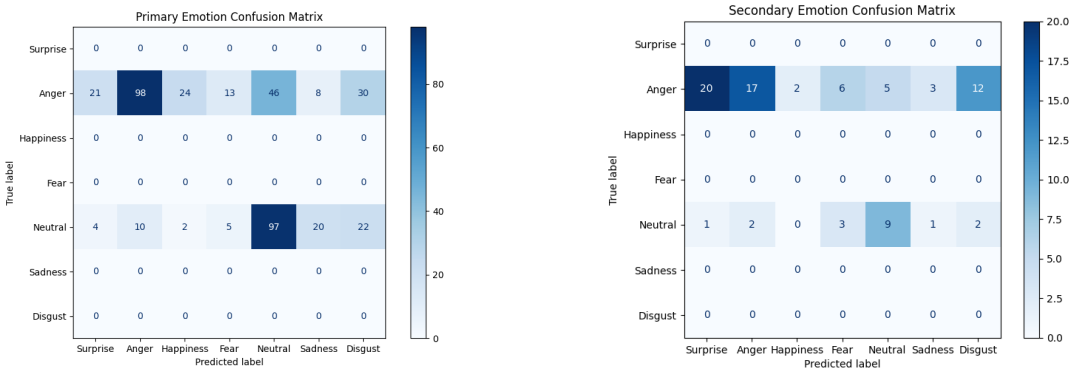

Figure 4: Confusion Matrix for Emotions judged as "Primary" (Left) and "Secondary" (Right)

Regarding the confusion matrices in Figure 4, we see that Surprise, Disgust, and the Neutral state were the most often misinterpreted emotions for Anger across primary and secondary choices. Contrarily, Sadness and Disgust were misinterpreted for the Neutral state. These results are quite interesting, as previous literature concerning the dimensional approach and acoustic-oriented approach would determine that Surprise tends to be quite far from Anger, while Disgust is phonetically lower than Anger. This carries over to the Neutral state as well, which is generally characterized by a "lack" of strong emotions, situating itself in the center of the circumplex model as shown in the Figure 3, while Disgust and Sadness lay at the extreme ends of valence, with only a difference of slight positive and negative arousal, respectfully.

There may be several reasons why the results for recognizing the base emotions turned out this way. Of course, we can acknowledge that the program was not able to accurately synthesize anger or that the changes in arousal and valence levels had them be more likely to be confused with other emotions, since indeed, we do see that lower intensity levels of anger were often confused with other emotions more than the other intensity levels. As the lowest intensity level of Anger was near the parameters that Neutral was at, there is little question as to why or how it had such a high confusion rate. Though this might be sufficient explanation for Surprise and Disgust as well, close analysis of each question showed no real relationship between intensity level and frequency of the chosen emotion, finding themselves as the most misinterpreted emotion across all levels of intensity. However, there are two other explanations as to why these results may have occurred. While only two emotional states were ever synthesized and presented, participants were able to choose from seven total emotional states, reflecting the basic emotions of Ekman and Cordaro (2011). Because of this, there is a non-zero chance that participants were looking for more emotions than there actually were. If there were less emotions to choose from, for instance deducing between Anger, Happiness, Neutral, and Sadness, we might have seen an increase of recognition.

Additionally, we can affirm that the context of the dialogue in the literature holds much importance. Lines 3 and 8, while coded as low and high intensity respectfully, resulted in 0 participants recognizing the emotion. Looking over the lines again, and putting the dialogue portion in isolation, we can see that the dialogue is quite ambiguous in what it is conveying, which is also represented in Figure 4, with Happiness scoring quite highly due to this contradiction in perceived sentiment and written sentiment. The lines in complete contain descriptors such as “said [...] angrily” or are contextually in a stressful event, resulting in the classifier to give them higher scores for anger. Yet, adding the context to the lines would skew the results, harming more than aiding it, as participants would be listening and/or reading the context of the line, rather than purely listening to the emotion behind it. Overall, despite the wide range of emotional choices, we can still affirm that the intended emotions were still recognized a majority of the time, and thus confirming the third hypothesis.

4.2 Intensity Level Recognition Analysis

As for intensity, scores were much poorer, reflecting a wide range of choices. To determine the accuracy, we looked at how many participants correctly recognized the intensity level out of the participants that correctly recognized the emotion. As the focus was Anger, and you cannot necessarily make Neutral more or less intense, intensity level results for neutral were discarded, Accuracy results were split into two points of reference, including and excluding the +1/-1 scale counts. Since we can deduce what intentions the participants had with the transitionary scales, they were included in the calculations. Among participants who recognized the emotion as “primary”, we achieved a 13.09% accuracy rate for correct intensity recognition. While incorporating the transitionary score increases the accuracy to 30.18%, this still indicates that intensity levels could not be consistently recognized by participants. When we add in “secondary emotion” responses, we achieve an increase of 16.56% and 32.65% respectfully, which does not change the overall conclusion.

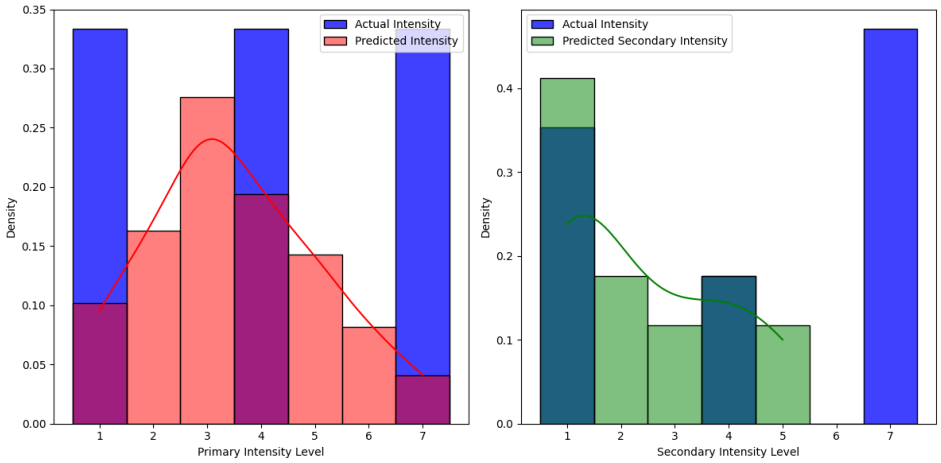

Figure 5: Histograms of Anger chosen as a ”Primary” emotion (Left) and ”Secondary” emotion (Right).

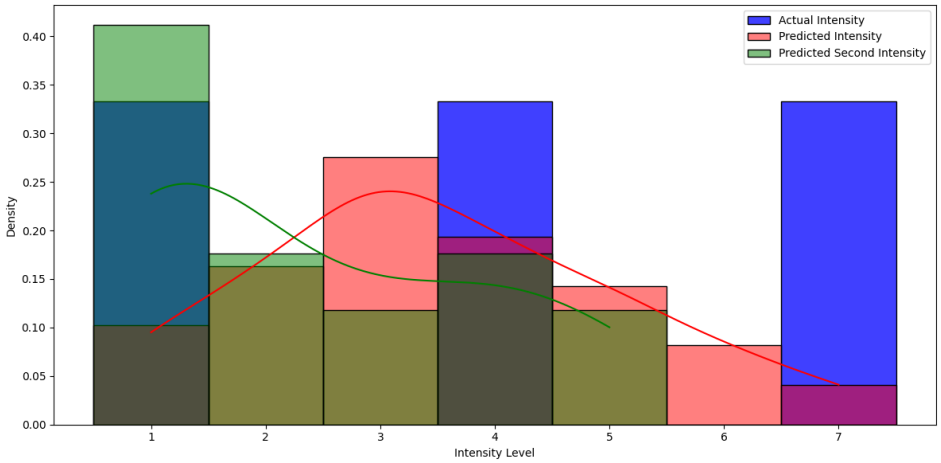

Figure 6: Combined Anger Intensity Level Histogram

The histograms of Figure 5 and 6 show interesting insights into the participant's thought processes. Figure 5 splits the data into primary and secondary emotion choices and builds the density off those responses. As not every participant believed or correctly chose a secondary emotion, the density of the secondary intensity (intensity levels chosen for Anger chosen as a secondary emotion) level chart varies from that of the primary intensity. From both charts, we see that there is a tendency to choose lower intensity levels amongst the participants, evident in a .27 density level for a primary intensity level of 3 and a .41 density level for secondary intensity level 1. The predicted primary intensity levels may indicate a hesitancy for choosing the extremes of 1 or 2, while the choosing a neutral or center level may be safer. This might be due to a lack of reference for how Anger at the different levels sounds like, forcing individuals to draw upon their own experiences, whether personally or through media, to make their decisions. One interesting result is the choice of secondary intensity level 1 that overtook the density of the actual intensity, which is also evident in the merged histograms in Figure 6, showing an overwhelming choice for the intensity level. Even taking into account the different weights of the answer, in which there were vastly less responses for secondary intensity levels than primary intensity levels, the score of 1 might reach that of the primary intensity score of 3 in how often it was chosen. One possible explanation to this occurrence could lie in the fact that these are "secondary" emotions, which may have caused the participants to believe that they must have lower intensities than the "primary" emotion of the audio.

These results may be explained by a couple reasons. The program may not have been able to create nuanced emotions, or the values chosen did not reflect the proper intensity, thus being unable to be correctly recognized. Furthermore, since this was a subjective study, and was based on individual's perceptions, rather than group or cooperative perceptions, the results are much more varied and reflect the differences in how individuals determine intensity levels as is evident in Figure 5 and Figure 6. Nevertheless, this indicates a falsification in two points of the hypothesis, mainly that this was not able to create a sufficient generalized basis for transitioning sentiment classification scores to arousal and valence values, and that individuals were not able to consistently recognize correct intensity levels in nuanced emotional synthesis.

4.3 Contributions

This study demonstrated the application of a FastSpeech2-based model for emotional speech synthesis, showing the use and feasibility of a quick and efficient method of synthesizing emotions. Furthermore, it highlighted that an older model, or one that is dependent on outdated modules, can still be reliable trained and perform its tasks effectively.

Despite displaying the challenges in creating recognizable nuanced expressive speech, this research creates a basis to further refine and address particular issues that limited the current model. It showed a better-than-chance result for recognizing "basic" emotions in nuanced speech, providing further groundwork for future improvements.

The falsification of the hypothesis emphasizes the importance of context for literature dialogues. It indicates that while dialogues in isolation may not be perfectly recognized, incorporating neutral renditions of sentiment descriptors alongside synthesized dialogues may help in progressing the naturalness of machine-generated expressive speech. Additionally, this study reveals the variability in human perception of nuanced emotions, especially when relying only on the dimensional features, illustrating the difficulties in creating a generalized method for synthesizing recognizable nuanced speech and transferring sentiment classification from text to emotional speech.

4.4 Limitations

While this study achieved some of its goals, there is still room for discussing how to improve it to either achieve better results or avoid consistent issues that plagued the study as it was progressing.

The first point of contention is the model itself. While Expressive-FastSpeech2 allowed for ease of implementation and detailed variability, and a faster training time characteristic of FastSpeech2, several issues provided difficulty to the training and inference process. The repository is dependent on older version of modules, many of which having features that have been depreciated or merged into a different function over the years. While finding workarounds was simple, this created difficulties in implementing a completely autonomous transition between the emotion classification model and Expressive-FastSpeech2. Furthermore, several issues were discovered after fully training the model, and due to the lack of time to retrain and fix the issues, they were left in. This included possible misalignment of MFA, which may have contributed to voices being used that were not indicative of the chosen speaker parameter, or the incorrect emotion being synthesized, despite other texts retaining the same parameters, though this may be due to the construction of the particular text. Furthermore, while a universal HiFi-GAN vocoder was used, because it was not specifically trained for the IEMOCAP dataset, there was a lot more noise present within the resulting synthesized speech, as well as cutting off the end of the provided text. If this research was to be done again, reviewing and changing the older dependencies to newer versions, as well as fine-tuning the vocoder would allow for a smoother synthesis and the autonomous function that was thought of at the beginning of the study.

Another such point was the requirement for human interference, as the emotional classifier had a subjective accuracy rate of 75%, forcing the researcher to determine which texts were true in their classification. This carries over the need to physically check the alignment of Expressive-FastSpeech2, and fix any mistakes done by MFA by hand.

While the speaker chosen sufficed for the experiment, it would have been better to spend more time analyzing each speaker choice, in order to determine which would result in the best audio for the study. In future recreations of this study, Tacotron2 could be trained on the same dataset with similar parameters to be compared to Expressive-FastSpeech2, as its accuracy in synthesis is near equal to FastSpeech2, and the majority of expressive synthesis models use Tacotron2 as the base architecture.

Finally, regarding the survey, there were a couple points that were brought up by the participants. Due to the poor quality of the synthesized text, even after being put through a filter, participants stated that it was still hard to ignore, even after being instructed to, and that it interfered with their ability to hear the emotion at times. While this was a known issue, the steps taken to avoid such problems were the extent of what could be done due to time restrictions. Furthermore, because of the quality, some participants stated that they had to force themselves to not infer the emotion from the provided transcription of the dialogue. While this is another valid concern, the questionnaire was asked in a way so that inferring from the text would only allot minimal information. One way to circumvent this issue entirely would be to completely get rid of the transcription of the audio, forcing participants to only listen to the audio. Though, if the same quality of audio is to be used, it would persistently cause difficulties in determining the emotion. Future studies may wish to focus on improving the audio quality in a way that it does not detract from the emotional parameters.

5 Conclusion

In summary, this thesis attempted to address a novel way of translating sentiment analysis of text to emotional nuanced speech, with a focus in angry and neutral emotional states.

Using an emotional text classifier model based on a Distilled version of RoBERTa (Hartmann, 2022), we classified every line of Alice in Wonderland (Carroll, 2006) into each one of Ekman and Cordaro (2011) basic emotions, including a neutral state, and refined the lists further until we were left with twenty lines of dialogue - twelve that represented three different intensity levels of anger based on probability scores, and eight that represented neutral states.

Afterwards, we trained an Expressive-FastSpeech2 model (K. Lee, 2021) on the IEMOCAP database (Busso et al., 2008). After pilot testing to determine the best values to use for the arousal and valence parameters for each level of anger, we synthesized all of our chosen lines of dialogue with their respective parameters. Due to a large amount of noise affecting the quality, we passed all the samples through noise filters created by Adobe (2022) or VoiceFixer(H. Liu et al., 2021).

We asked twenty participants to evaluate the emotion and intensity in each of the synthesized samples, also allowing them to choose a second emotion and subsequent intensity if they believed there was another emotion present.

5.1 Summary of Findings and Contributions

Our results revealed an above random chance accuracy rate of 55.25% for participants recognizing the basic emotion in the audio when incorporating responses that answered it as a secondary emotion, and a low accuracy of 32.65% for recognizing the intensity level when adding secondary emotion intensities and answers that fell within 1 stage of the correct level.

The results partially proved the third hypothesis, in such that the coded emotions could be consistently recognized in the nuanced synthesis, at a rate of 55.25%, but disproved the first and second hypotheses, indicating that recognizing and synthesizing intensity levels still require further research, and that a generalized model for transforming sentiment scores to arousal and valence values continues to face challenges.

Although the hypothesis was falsified, several notes of interest were still discovered and emphasized from the results. These insights include the importance of descriptive context in novels when synthesizing dialogues, and the variability of human perception of nuanced emotions, both of which are areas of study that once addressed, can lead to future progress in machine emotional intelligence and human-machine interaction.

In the scope of audiobooks, this study still holds promise in creating a generalized reference for interpretation. While nuanced speech still requires work, we created a quick and efficient process from this combination of models that reflect the intended basic emotions of a piece of literature, with an accuracy in recognition of above random chance. Despite the requirement of human subjectivity to better refine the emotion of the dialogue on their own, the possible reduction in time spent is promising.

5.2 Future Work

In light of these findings, this thesis opens up several paths of future study. The first option is to attempt this study once more on the same model, but fix the issues that caused the low-quality syn-

thesis, such as training the vocoder on the dataset and double-checking the MFA alignment manually. However, as an extension, improving the Expressive-FastSpeech2 model to reduce the amount of human interference or upgrading the older dependencies to new ones is another avenue one can take to make the training and synthesizing process more efficient. In updating the Expressive-FastSpeech2 model, this could also allow for the creation of an automatic process, in which a piece of literature is put into the program, which categorizes each line of dialogue or exposition into emotions, to then synthesize each line based on its emotion and probability score.

Another such way to recreate this study would be to test it on a different model, such as one with a Tacotron2-based architecture. As mentioned in Section 4.4, Tacotron2 shares a similar accuracy in synthesis with FastSpeech2 and has been used by a majority of the more recent expressive-speech models as a basis.

This study could also explore other emotions and their own intensity levels, comparing with one another to see how often they are recognized or confused.

Finally, the survey could be re-conducted in a couple of different ways. The first method would be to reduce the number of emotions participants could answer from, removing further chances of confusion. Additionally, if only focusing on intensity levels, one could provide human-acted references to participants so they knew what baselines of each level they were looking for, which would allow for responses based off a common reference.

References

A., F. S., V.R., V. K., A., R. S., Jayakumar, A., & P., B. A. (2009). Speaker independent automatic emotion recognition from speech: A comparison of mfccs and discrete wavelet transforms. *2009 International Conference on Advances in Recent Technologies in Communication and Computing*. doi: 10.1109/artcom.2009.231

Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020, 05). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, *2*. doi: 10.1002/eng2.12189

Adobe. (2022). *Enhance speech from adobe — free ai filter for cleaning up spoken audio.* Author. Retrieved 2024-05-23, from `https://podcast.adobe.com/enhance#`

Bharti, S. K., Varadhaganapathy, S., Gupta, R. K., Shukla, P. K., Bouye, M., Hingaa, S. K., & Mahmoud, A. (2022, 08). Text-based emotion recognition using deep learning approach. *Computational Intelligence and Neuroscience*, *2022*, 1-8. doi: 10.1155/2022/2645381

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., ... Narayanan, S. S. (2008, 11). Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*, 335-359. doi: 10.1007/s10579-008-9076-6

Carroll, L. (2006). *Alice's adventures in wonderland*. Sam'l Gabriel Sons & Company.

Christie, I. C., & Friedman, B. H. (2004, 01). Autonomic specificity of discrete emotion and dimensions of affective space: a multivariate approach. *International Journal of Psychophysiology*, *51*, 143-153. doi: 10.1016/j.ijpsycho.2003.08.002

Daly, E. M., Lancee, W. J., & Polivy, J. (1983). A conical model for the taxonomy of emotional experience. *Journal of Personality and Social Psychology*, *45*, 443-457. doi: 10.1037/0022-3514.45.2.443

Devlin, J., Chang, M.-W., Lee, K., Google, K., & Language, A. (2019, 05). *Bert: Pre-training of deep bidirectional transformers for language understanding.*

Ekman, P., & Cordaro, D. (2011, 09). What is meant by calling emotions basic. *Emotion Review*, *3*, 364-370. doi: 10.1177/1754073911410740

Guerrero, L., Andersen, P., & Trost, M. (1996, 01). *Communication and emotion: Basic concepts and approaches.* Academic Press.

Gunes, H., Schuller, B., Pantic, M., & Cowie, R. (2011, 03). *Emotion representation, analysis and synthesis in continuous space: A survey.* doi: 10.1109/FG.2011.5771357

Harmon-Jones, E., Harmon-Jones, C., & Summerell, E. (2017, 09). On the importance of both dimensional and discrete models of emotion. *Behavioral Sciences*, *7*, 66. doi: 10.3390/bs7040066

Hartmann, J. (2022). *Emotion english distilroberta-base.* Retrieved 2024-05, from `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/`

Hoffmann, H., Scheck, A., Schuster, T., Walter, S., Limbrecht, K., Traue, H., & Kessler, H. (2012, 02). *Mapping discrete emotions into the dimensional space: An empirical approach.* 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC). doi: 10.1109/ICSMC.2012.6378303

Izard, C. E. (2011, 09). Forms and functions of emotions: Matters of emotion–cognition interactions. *Emotion Review*, *3*, 371-378. doi: 10.1177/1754073911410737

Kloves, S., & Audsley, M. (Eds.). (2005, 11). *Harry potter and the goblet of fire.* Warner Bros. Pictures.

Kolita, S., & Acharjee, P. (2021). Evaluation of the impact of emotion in assamese emotional speech

signals. *International Journal of Mechanical Engineering*, *6*, 974-5823.

Laukka, P., Juslin, P., & Bresin, R. (2005, 08). A dimensional approach to vocal expression of emotion. *Cognition & Emotion*, *19*, 633-653. doi: 10.1080/02699930441000445

Lee, K. (2021). *Expressive-fastspeech2.* Github. Retrieved 2024-04, from `https://github.com/keonlee9420/Expressive-FastSpeech2`

Lee, Y., Rabiee, A., & Lee, S.-Y. (2017, 11). *Emotional end-to-end neural speech synthesizer.*

Lei, Y., Yang, S., Wang, X., & Xie, L. (2022). Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 853-864. doi: 10.1109/taslp.2022.3145293

Levenson, R. W. (2011, 09). Basic emotion questions. *Emotion Review*, *3*, 379-386. doi: 10.1177/1754073911410743

Li, M. (2023). *Fine-tuned distilroberta-base for emotion classification.* Retrieved 2024-05, from `https://huggingface.co/michellejieli/emotion_text_classifier`

Liu, H., Kong, Q., Tian, Q., Zhao, Y., Wang, D., Huang, C., & Wang, Y. (2021). *Voicefixer: Toward general speech restoration with neural vocoder.* GitHub. Retrieved 2024-05-23, from `https://github.com/haoheliu/voicefixer`

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Allen, P. (2019). *Roberta: A robustly optimized bert pretraining approach.*

Luo, X., Takamichi, S., Saito, Y., Koriyama, T., & Saruwatari, H. (2024). Emotion-controllable speech synthesis using emotion soft label, utterance-level prosodic factors, and word-level prominence. *APSIPA Transactions on Signal and Information Processing*, *13*. doi: 10.1561/116.00000242

Mohanta, A., & Mittal, V. (2017). *Human emotional states classification based upon changes in speech production features in vowel regions.*

Panksepp, J., & Watt, D. (2011, 09). What is basic about basic emotions? lasting lessons from affective neuroscience. *Emotion Review*, *3*, 387-396. doi: 10.1177/1754073911410741

Park, J.-S., Jang, G.-J., & Kim, J.-H. (2012, 08). Multistage utterance verification for keyword recognition-based online spoken content retrieval. *IEEE Transactions on Consumer Electronics*, *58*, 1000-1005. doi: 10.1109/tce.2012.6311348

Plutchik, R. (1980). *Emotion, a psychoevolutionary synthesis*. Harper & Row.

Plutchik, R., & Kellerman, H. (1980). *Emotion : theory, research and experience*. Academic Press.

Polzehl, T., Schmitt, A., Metze, F., & Wagner, M. (2011, 11). Anger recognition in speech using acoustic and linguistic cues. *Speech Communication*, *53*, 1198-1209. doi: 10.1016/j.specom.2011.05.002

Razak, A. A., Abidin, M. I. Z., & Komiya, R. (2003). *Emotion pitch variation analysis in malay and english voice samples.* IEEE. doi: 10.1109/apcc.2003.1274322

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liue, T.-Y. (2022, 08). Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv*. doi: 10.48550/arXiv.2006.04558

Rowling, J. (2000). *Harry potter and the goblet of fire*. Arthur A Levine.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161-1178. doi: 10.1037/h0077714

Santoro, A., & Marcelli, A. (2019, 09). A novel procedure to speed up the transcription of historical handwritten documents by interleaving keyword spotting and user validation. *2019 International Conference on Document Analysis and Recognition (ICDAR)*. doi: 10.1109/icdar.2019.00198

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... Wu, Y. (2018, 04). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi: 10.1109/icassp.2018.8461368

Sobin, C., & Alpert, M. (1999, 07). Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy. *Journal of Psycholinguistic Research*, *28*, 347-365.

Tracy, J. L., & Randles, D. (2011, 09). Four models of basic emotions: A review of ekman and cordaro, izard, levenson, and panksepp and watt. *Emotion Review*, *3*, 397-405. doi: 10.1177/1754073911410747

Zagalo, N., Torres, A., & Branco, V. (2005). Emotional spectrum developed by virtual storytelling. *Lecture Notes in Computer Science*, 105-114. doi: 10.1007/11590361_12

Zhou, K., Sisman, B., Rana, R., Schuller, B. W., & Li, H. (2023, 10). Speech synthesis with mixed emotions. *IEEE Transactions on Affective Computing*, *14*, 3120-3134. doi: 10.1109/taffc.2022.3233324

Appendices

A Lines and Emotion Probabilities

The following are the lines chosen and used for evaluation, along with their top 3 emotions and subsequent scores they were classified with. Sentences labeled as Neutral did not have their probability scored. They are in the same order as they were seen in the survey.

1. Come, there's no use in crying like that!" said Alice to herself rather sharply. "I advise you to leave off this minute!" She generally gave herself very good advice (though she very seldom followed it), and sometimes she scolded herself so severely as to bring tears into her eyes.
   Top Three Labels: ['anger', 'disgust', 'fear']
   Top Three Scores: [0.7572353482246399, 0.181224063038826, 0.022262414917349815]

2. This question the Dodo could not answer without a great deal of thought. At last it said, "Everybody has won, and all must have prizes."
   Label: neutral
   Score: 1.0

3. She waited for some time without hearing anything more. At last came a rumbling of little cart-wheels and the sound of a good many voices all talking together. She made out the words: "Where's the other ladder? Bill's got the other–Bill! Here, Bill! Will the roof bear?–Who's to go down the chimney?–Nay, I sha'n't! You do it! Here, Bill! The master says you've got to go down the chimney!"
   Top Three Labels: ['anger', 'surprise', 'fear']
   Top Three Scores: [0.5424417853355408, 0.14933885633945465, 0.09269830584526062]

4. "I do," Alice hastily replied; "at least–at least I mean what I say–that's the same thing, you know."
   Label: neutral
   Score: 1.0

5. "Speak English!" said the Eaglet. "I don't know the meaning of half those long words, and, what's more, I don't believe you do either!"
   Top Three Labels: ['anger', 'disgust', 'surprise']
   Top Three Scores: [0.5338663458824158, 0.10816691070795059, 0.2613014876842499]

6. "I know what 'it' means well enough, when I find a thing," said the Duck; "it's generally a frog or a worm. The question is, what did the archbishop find?"
   Label: neutral
   Score: 1.0

7. "If there's no meaning in it," said the King, "that saves a world of trouble, you know, as we needn't try to find any. Let the jury consider their verdict."
   Label: neutral
   Score: 1.0

8. ”It is a very good height indeed!” said the Caterpillar angrily, rearing itself upright as it spoke (it was exactly three inches high).
Top Three Labels: [’anger’, ’disgust’, ’neutral’]
Top Three Scores: [0.9553799033164978, 0.01285378448665142, 0.0189078189432621]

9. ”No, no!” said the Queen. ”Sentence first–verdict afterwards.”
Top Three Labels: [’anger’, ’disgust’, ’fear’]
Top Three Scores: [0.7322607636451721, 0.03360165283083916, 0.19622720777988434]

10. The judge, by the way, was the King and he wore his crown over his great wig. ”That’s the jury-box,” thought Alice; ”and those twelve creatures (some were animals and some were birds) I suppose they are the jurors.”
Label: neutral
Score: 1.0

11. Soon her eye fell on a little glass box that was lying under the table: she opened it and found in it a very small cake, on which the words ”EAT ME” were beautifully marked in currants. ”Well, I’ll eat it,” said Alice, ”and if it makes me grow larger, I can reach the key; and if it makes me grow smaller, I can creep under the door: so either way I’ll get into the garden, and I don’t care which happens!”
Top Three Labels: [’anger’, ’disgust’, ’fear’]
Top Three Scores: [0.5422682762145996, 0.24589607119560242, 0.10008420795202255]

12. ”Well, it’s got no business there, at any rate; go and take it away!”
Top Three Labels: [’anger’, ’disgust’, ’fear’]
Top Three Scores: [0.949924647808075, 0.014087660238146782, 0.022689953446388245]

13. Very soon the Rabbit noticed Alice, and called to her, in an angry tone, ”Why, Mary Ann, what are you doing out here? Run home this moment and fetch me a pair of gloves and a fan! Quick, now!”
Top Three Labels: [’anger’, ’neutral’, ’surprise’]
Top Three Scores: [0.9226188063621521, 0.017316516488790512, 0.025453316047787666]

14. ”What do you mean by that?” said the Caterpillar, sternly. ”Explain yourself!”
Top Three Labels: [’anger’, ’disgust’, ’fear’]
Top Three Scores: [0.7912983298301697, 0.04230301082134247, 0.10509485751390457]

15. ”Who cares for you?” said Alice (she had grown to her full size by this time). ”You’re nothing but a pack of cards!”
Top Three Labels: [’anger’, ’disgust’, ’neutral’]
Top Three Scores: [0.5947501063346863, 0.18572679162025452, 0.10066939145326614]

16. The Hatter opened his eyes very wide on hearing this, but all he said was ”Why is a raven like a writing-desk?”
Label: neutral
Score: 1.0

17. ”Not like cats!” cried the Mouse in a shrill, passionate voice. ”Would you like cats, if you were me?”
Top Three Labels: [’anger’, ’disgust’, ’fear’]
Top Three Scores: [0.7547333240509033, 0.06565763801336288, 0.09514901041984558]

18. ”You insult me by talking such nonsense!” said the Mouse, getting up and walking away.
Top Three Labels: [’anger’, ’disgust’, ’fear’]
Top Three Scores: [0.9495113492012024, 0.027451995760202408, 0.007980800233781338]

19. ”You mean you can’t take less,” said the Hatter; ”it’s very easy to take more than nothing.”
Label: neutral
Score: 1.0

20. ”You might just as well say,” added the Dormouse, which seemed to be talking in its sleep, ”that ’I breathe when I sleep’ is the same thing as ’I sleep when I breathe!’”
Label: neutral
Score: 1.0

B Survey Questions

The following is an example of how the survey questions were conducted. All the questions were the same, save for the audio and transcription.

Start of Block: Audio 1

-Audio Sample-

”Come, there’s no use in crying like that! I advise you to leave off this minute!”

Q1 What is the primary emotion in the audio sample?

o Happiness (1)
o Sadness (2)
o Anger (3)
o Neutral (4)
o Disgust (5)
o Fear (6)
o Surprise (7)

Q2 What is the level of intensity of the emotion portrayed in the audio sample? (1 - Low Intensity ; 7 - High Intensity)

o 1 (1)
o 2 (2)
o 3 (3)
o 4 (4)
o 5 (5)
o 6 (6)
o 7 (7)

Q3 Is there another emotion recognizable in the audio sample?

o No (1)
o Yes (2)

Q4 What is the second most prevalent emotion in the audio sample?

o Happiness (1)
o Sadness (2)
o Anger (3)
o Neutral (4)
o Surprise (5)
o Disgust (6)
o Fear (7)

Q5 What is the level of intensity of the secondary emotion portrayed in the audio sample? (1 - Low Intensity ; 7 - High Intensity)

o 1 (1)
o 2 (2)
o 3 (3)
o 4 (4)
o 5 (5)
o 6 (6)
o 7 (7)

End of Block: Audio 1

C Consent Form

The following is the consent form participants were provided at the beginning of the survey. If they did not consent to the collection of the data and study, the survey was automatically ended for them and they could not fill out the rest of the survey.

Thank you for your interest and participation in this study!

Background The purpose of this study is to investigate to what extent emotions can be conveyed through Text-to-Speech synthesized audio. To explore this, we have created a questionnaire that contains audio files which portray emotions at different levels of intensity. You will be asked to judge which is the primary emotion from a selection of 6 basic emotions, in addition to a neutral state, and at what intensity the primary emotion is at. If you believe there are multiple emotions present, you will have the opportunity to add another emotion and its intensity level to your answer.
Confidentiality Your responses will be kept confidential. Data collected will be saved anonymously and used solely for research purposes. No personally identifiable information will be asked or linked to your responses. Participation in this study is entirely voluntary.
Consent By doing this online survey and submitting my response, I agree to the following:

- I have read and understood the project information above
- I understand that taking part is voluntary
- I understand that the data I provide is saved anonymously and is accessible by members of the research team.
- I agree that the data that I provide is used for research purposes and research dissemination

This study is being performed by Jocomin Galarneau of the MSc Voice Technology of the University of Groningen. If you have any questions, please contact the researcher via the email: j.t.l.m.galarneau@student.rug.nl
o I consent (1)
o I do not consent (2)